THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

September 24, 2004

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.

APPLICATION NUMBER: *60/478,780*
FILING DATE: *June 16, 2003*

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

By Authority of the
COMMISSIONER OF PATENTS AND TRADEMARKS

M. SIAS
Certifying Officer

06-17-03

PTO/SB/16 (10-01)
Approved for use through10/31/2002. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(c).

Express Mail Label No. EF195547993US

INVENTOR(S)

Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
Herwig, Gaston, Emiel	VAN MARCK	Nazareth, Belgium
Tim, Gerrit	VAN DEN BULCKE	Mechelen, Belgium

□ *Additional inventors are being named on the ______ separately numbered sheets attached hereto*

TITLE OF THE INVENTION (500 characters max)

QUANTITATIVE PREDICTION METHOD

Direct all correspondence to: **CORRESPONDENCE ADDRESS**

☒ Customer Number 000027777
Type Customer Number here → *Place Customer Number Bar Code Label here*

OR

□ Firm *or* Individual Name					
Address					
Address					
City		State		ZIP	
Country		Telephone		Fax	

ENCLOSED APPLICATION PARTS *(check all that apply)*

☒ Specification *Number of Pages*	35	□ CD(s), Number	
☒ Drawing(s) *Number of Sheets*	16	□ Other (specify)	
□ Application Data Sheet. See 37 CFR 1.76			

METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT

□ Applicant claims small entity status. See 37 CFR 1.27.
□ A check or money order is enclosed to cover the filing fees
☒ The Commissioner is hereby authorized to charge filing fees or credit any overpayment to Deposit Account Number: 10-0750/VIP0022/JJIT
□ Payment by credit card. Form PTO-2038 is attached.

FILING FEE AMOUNT ($) 160.00

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.
☒ No.
□ Yes, the name of the U.S. Government agency and the Government contract number are: __________

Respectfully submitted,

SIGNATURE __________ Date 16 JUNE 2003

TYPED or PRINTED NAME Jesús Juanós i Timoneda

REGISTRATION NO. *(if appropriate)* 43,332

TELEPHONE (732) 524-1513

Docket Number: VIP 0022

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

This collection of information is required by 37 CFR 1.51. The information is used by the public to file (and by the PTO to process) a provisional application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 8 hours to complete, including gathering, preparing, and submitting the complete provisional application to the PTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, Washington, D.C. 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Box Provisional Application, Assistant Commissioner for Patents, Washington, D.C. 20231.

QUANTITATIVE PREDICTION METHOD

The present invention concerns methods and systems for analysis of drug resistance in HIV-1. More specifically, the invention provides methods for predicting drug resistance by correlating genotypic information with phenotypic profiles. The methods allow the identification of primary and secondary resistance-associated mutations for new and existing drugs and for calculating the contribution of mutations and combinations of mutations to resistance and hypersusceptibility. The invention allows the design, optimization and assessment of the efficiency of a therapeutic regimen based upon the genotype of the disease affecting a patient.

All publications, patents and patent applications cited herein are incorporated in full by reference.

BACKGROUND

Techniques to determine the resistance of HIV-1 to a therapeutic agent are becoming increasingly important. Many patients experience treatment failure or reduced efficacy over time. This is generally due to the virus mutating and/or developing a resistance to the treatment. As used herein, "HIV" is the human immunodeficiency virus, which is a retrovirus.

The various different anti-HIV-1 agents that have been developed over the years were initially administered to patients alone, as monotherapy. Though a temporary antiviral effect was observed, all the compounds lost their effectiveness over time. Research has now demonstrated that one of the main reasons behind treatment failure for all the antiviral drugs is the development of resistance of the virus to the drug (see, for example, Larder et al., 1989, Science, 246, 1155-8). This is largely due to the ability of HIV continuously to generate a number of genetic variants in a replicating viral population. These genetic changes generally alter the configuration of the HIV reverse transcriptase (RT) and protease (PR) molecules in such a way that they are no longer susceptible to inhibition by compounds developed to target them. If antiretroviral therapy is ongoing and if viral replication is not completely suppressed, the selection of genetic variants is inevitable and the viral population becomes resistant to the drug.

Since then, dual combination therapy, using drugs that target both HIV reverse transcriptase (RT) and protease (PR) molecules, has provided increased control of viral replication, and thus provided extended clinical benefit to patients. In recent years, however, it has become clear that even patients being treated with triple therapy including a protease inhibitor often eventually experience treatment failure.

Since patients in the developed world are generally prescribed cocktails of therapeutic drugs, not all HIV-1 infections originate with a wild type, drug sensitive strain from which drug resistance will emerge - with the increase in prevalence of drug resistant strains comes the increase in infections that actually begin with drug resistant strains. Infections with pre-existing drug resistance immediately reduce the drug options for drug treatment and emphasize the importance of drug resistance information to optimize initial therapy for these patients.

Moreover, as the number of available antiretroviral agents has increased, so has the number of possible drug combinations and combination therapies. It is therefore very difficult, if not impossible, for the physician to establish the optimal combination for an individual. Although there are many drugs available for use in combination therapy, the choices can quickly be exhausted and the patient can rapidly experience clinical progression or deterioration if the wrong treatment decisions are made. The key to tailored, individualized therapy lies in the effective profiling of the individual patient's virus population in terms of sensitivity or resistance to the available drugs. This requires the advent of truly individualized therapy.

There are certain solutions to this problem currently in use.

Phenotyping directly measures the actual sensitivity of a patient's pathogen or malignant cell to particular therapeutic agents. However, this can be slow, labour-intensive and thus expensive.

A second approach to measuring resistance involves genotyping tests that detect specific genetic changes (mutations) in the viral genome which lead to amino acid changes in at least one of the viral proteins, known or suspected to be associated with resistance. Although genotyping tests can be performed more rapidly, a problem with genotyping is that there are now over 100 individual mutations with evidence of an effect on susceptibility to HIV-1 drugs and new ones are constantly being discovered, in parallel with the development of new drugs and treatment strategies. The relationship between these point mutations, deletions and insertions and the actual susceptibility of the virus to drug therapy is extremely complex and interactive. An example of this complexity is the M184V mutation that confers resistance to 3TC but reverses AZT resistance. The 333D/E mutation, however, reverses this effect and can lead to dual AZT/3TC resistance.

Sophisticated interpretation is therefore required to predict what the net effect of these mutations might be on the susceptibility of the virus population to the various therapeutic agents. Rules-based computer algorithms have provided some assistance, for example, see International patent application WO01/79540. An overview of this type of technique is presented in Figure 1: (figure 1 from poster). However, there remains a continuing need for the quantitative prediction of HIV drug susceptibility from viral genotype. Furthermore, because the majority of HIV patients have now been exposed to drug cocktails, it is thought that the disease-causing retroviruses tend to spontaneously generate mutations that have often co-evolved. This makes the analysis of which mutations are responsible for resistance to which drugs almost impossible using currently available techniques. It also means that mutations that contribute to resistance are being overlooked using the currently available analysis techniques.

It is therefore an aim of the present invention to provide methods for improving the interpretation of genotypic results.

It is a further aim of the invention to provide methods for determining (or predicting) a phenotype based on a genotype.

It is also a further aim of the invention to provide methods for predicting the resistance of an HIV variant of a particular genotype to a therapy or a therapeutic agent.

It is also an aim of the invention to predict resistance of a patient to therapy.

It is also an aim of the invention to provide methods to assess the effectiveness or efficiency of a therapy or to optimize a patient's therapy.

It is also an aim of the invention to identify novel HIV-1 mutations that are associated with resistance to particular drug therapies or combination therapies.

SUMMARY OF THE INVENTION

A solution to these problems involves new methods for measuring drug resistance by correlating genotypic information with phenotypic drug resistance profiles measured experimentally.

According to a first aspect of the invention, there is provided a method for quantitating the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by HIV, said method comprising the steps of:

a) performing a linear regression analysis using data from a dataset of matching genotypes and phenotypes, whereby the log fold resistance, pFR, is modelled as the sum of all the individual resistance contributions for each of the mutations or combinations of mutations that occur in HIV according to the following equation;

$$pFR = \beta_A M_A + \beta_B M_B + \cdots + \beta_Z M_Z + \varepsilon$$

wherein each individual resistance contribution is calculated by multiplying a mutation factor, M_A , M_B , ..., M_Z, for each mutation or combination of mutations by a resistance coefficient β_A, β_B, ..., β_Z;

wherein the mutation factor assigned to each mutation or combination of mutations reflects the degree to which that mutation or combination of mutations is present in the HIV strain and, if present, to which degree the mutation is present in a mixture;

wherein each resistance coefficient reflects the contribution of the mutation or combination of mutations to the fold resistance exhibited by the strain;

and wherein the error term ε, represents the difference between the modelled prediction and the experimentally determined measurement.

This method involves a data driven technique for quantitative drug susceptibility prediction. This method uses a multiple linear regression model to estimate coefficient values that accurately reflect the contribution made by a particular HIV mutation or combination of mutations to resistance to a particular drug. Repeating the method for each candidate therapeutic drug allows the compilation of a global picture of drug resistance exhibited by a particular HIV strain.

This method has allowed the identification of mutations hitherto unrecognised as having an effect on drug resistance in HIV. The method also allows the identification of primary (single mutations) and secondary (the co-occurrence of two mutations) or higher order terms resistance-associated mutations for new and existing drugs. Accordingly, a further aspect of the invention provides a method of identifying a mutation that affects the degree of drug resistance exhibited by an HIV strain using a method according to the first aspect of the invention.

The method of the first aspect of the invention is also advantageous over current methods since it allows the quantitative, purely data-driven, objective_assessment of the contribution of mutations and combinations of mutations to drug resistance. The method also allows the deconvolution of the individual contribution made by particular mutations to the drug resistance phenotype. Unlike existing methods, the method is able to correct for correlating mutations that on the face of it appear to affect drug resistance, but which in fact only correlate in their occurrence with resistance causing mutations and are themselves phenotypically silent.

The method has allowed the design of an automated computational technique for the prediction of the drug resistance profile possessed by a particular HIV strain infecting a patient. The methods thus allow the determination of a patient phenotype without having to perform any phenotypic testing whatsoever. This has clear ramifications for the bespoke design, optimization and assessment of strategies for individual patient therapy based upon the genotype of the infecting agent.

The invention also provides diagnostic kits for performing each of the methods of the invention described herein.

In any population of HIV variants, there is a wide distribution of drug resistance phenotypes for any particular drug, ranging from hyper-susceptibility to strong resistance (see Figure 2 (figure 2 from poster)). The expression "drug resistance phenotype" means the resistance of an HIV virus to a tested therapy, therapeutic agent or drug. The term "resistance" as used herein, pertains to the capacity of resistance, sensitivity, susceptibility, or effectiveness of a therapy against a disease. The term "therapy" includes but is not limited to a drug, pharmaceutical, or any other compound or combination of compounds that can be used in therapy or therapeutic treatment of HIV. This distribution of drug resistance reflects the large number of different genotypes that are present in the population. Some variants may only have one mutation that is correlated with drug resistance, whilst others will have several or numerous such mutations, each of which may impart its own contribution to the drug resistance phenotype.

Adding an additional level of complication are the phenomena of antagonism, synergy and enhancement, where certain mutations may add to or detract from the effect of other mutations in a manner not predictable from studying the effects of the individual mutations alone. Highly correlated mutations are also problematic. These are mutations that almost always co-occur in a strain, but only one of the mutations actually has an effect on drug resistance. For example, when one of these 2 mutations has an effect on resistance and the other mutation does not (this mutation might for example be highly correlated with the resistance mutation because it affects the replication rate of the virus), the effect can erroneously be assigned to either one of the mutations.

Examples of mutations known or suspected to influence the sensitivity of HIV to drug therapy may be found on the internet at http://hiv-web.lanl.gov; http://hivdb.stanford.edu/hiv/; or http://www.viral-resistance.com.

In HIV, two sections of the genome are generally studied: Protease (PR) and Reverse Transcriptase (RT). The methods of the present invention can equally be applied to other sections of the HIV genome such as integrase (IN). A mutation is presented as a number referring to the position in the protein, followed by the amino acid(s) on that position, if it differs from the amino acid in the HXB2 HIV reference. In the terms included above, the mutations are represented as "A", "B", ..."Z".

Mixtures reflect the diversity of the HIV population in a sample. It means that on that position two subsets of the population have a different amino acid. Mixtures are denoted by separating amino acids with the '/' character: 65K/R (mixture of 'K' and 'R' at position 65).

When more than two amino acids are found on a certain position in subsets of the population, the dummy amino acid 'X' is used.

Insertions are denoted by adding the insert position behind a dot: 69.2S (an insert of 'S' at insert position 2). Deletions are denoted by a minus sign: 69-.

Examples of mutations present in the RT domain of HIV conferring resistance to a reverse transcriptase inhibitor include 69C, 69V, 69T, 75A, 1011, 103T, 103N, 184T, 188H, 190E, 219N, 219Q, 221Y, 2211, and 233V. Additional examples of mutations present in the protease (PR) domain of HIV conferring resistance to a reverse transcriptase inhibitor include 24M, 48A, and 53L. A mutation may affect resistance alone or in combination with other mutations.

For the purposes of the invention, the mutations identified should be associated with resistance or susceptibility to drug therapy, for example an antiretroviral drug. The degree to which a particular mutation pattern may affect resistance may be determined by one of skill in the art, for example, using the phenotypic resistance monitoring assay such as, the ANTIVIROGRAM® (Virco, Belgium) (see WO97/27480). In this methodology, resistance is determined with respect to a laboratory reference strain HIVLAI/IIIB. The difference in IC_{50} (the concentration of drug required to reduce the virus' growth in cell culture by 50%) between the patient sample and the reference viral strain is determined as a quotient. This fold change in IC_{50} is reported and indicative of the resistance profile of a certain drug. Based on the changes in IC_{50}, cut-off values have been established to distinguish a sample from being sensitive or resistant to a certain drug.

Various projects are underway to compile data relating to the correspondence of certain mutations with drug resistance phenotype, and these generally lead to the generation of relational databases of tables that illustrate the matching genotype / resistance phenotype for various antiretroviral drugs. Such databases bring together the knowledge of both a genotypic and phenotypic database. The phenotypic database contains phenotypic resistance values for HIV to at least one therapy, preferably multiple drug therapies. For example, the phenotypic resistance values of tested HIV viruses, with a fold resistance determination compared to the reference HIV virus (wild type).

The dataset used herein is a dataset developed by the Applicant, which consists of a set of matching genotype / phenotype measurements with possible multiple phenotype measurements per genotype. However, any similar dataset may be used, provided that there are sufficient entries for each genotype / phenotype measurement for the data to be significant. In the Virco dataset, the mutations are defined relative to HXB2 at amino acid level.

The phenotypes are presented as *pFR* values, where *pFR* is equal to - *log (FR)*, where *FR* denotes the Fold Resistance. Negative *pFR* values thus denote resistance and positive values denote hyper-susceptibility. For example, a *pFR* value of -1.0 is equal to 10-fold resistance. An example of the pFR distribution for Saquinavir (SQV) is shown in Figure 3 (slide 5). Figure 4 (slides 7 and 8) shows the pFR distribution for the "48V" mutation on SQV. It is clear from this that the 48V subset does not behave the same as the whole dataset.

The problems of unwanted correlations between mutations where not all correlated mutations contribute to the drug resistance phenotype are illustrated in Figure 5 (slide 9). Here, the left hand panel shows the pFR distribution for the 71I mutation. When the effects of mutations 48V and 84V are removed (right hand panel), the pFR distribution is markedly increased (less drug resistance).

According to the invention, the predicted fold resistance of an HIV strain of a particular genotype may be calculated by summing the individual resistance contributions for each of the mutations or combinations of mutations in the mutation pattern of that genotype. The method uses linear regression models, so that the phenotype prediction, *pFR* is calculated in the following equation (1):

$$pFR = \beta_A M_A + \beta_B M_B + \cdots + \beta_Z M_Z + \varepsilon$$

The independent variables M_A , M_B , ..., M_Z, are referred to herein as mutation factors, each of which reflects the degree to which the mutation or combination of mutations is present in the HIV strain and, if present, whether or not the mutation is present in a mixture.

The resistance coefficients β_A, β_B, ..., β_Z represent the contribution to the total *pFR* prediction for each single mutation.

Each mutation factor M_i thus represents the presence or absence of the corresponding mutation and each coefficient β_i represents the contribution to the *pFR* change for that specific mutation.

The mutation factor may take into account 1st order terms (single mutations) as well as 2nd order terms (the co-occurrence of two mutations) and in general nth order terms. For example, 2nd order terms take the form:

$$\beta_{AB} M_{AB}$$

The independent variable M_{AB} represents the co-occurrence of mutations A and B and the coefficient β_{AB} represents the synergy or antagonism between mutations A and B.

Higher order terms affect for interactions between mutations:

- reversal or antagonism: positive *pFR* shifts for mutation couples.
- synergy or enhancement: extra negative *pFR* shift for mutation couple.

For example, consider the following (artificial) model:

Mutation	Coefficient
A	-0.46
B	-0.92
C	-0.64
D	0.63
E	-0.16
F	-0.19
F & A	-0.09

Consider a virus with following mutations: F, A and E. Applying equation (1), this virus will have a *pFR* prediction:

$$pFR = \beta_F.1 + \beta_A.1 + \beta_E.1 + \beta_{A;F}.1 = -0.9$$

or almost 8-fold resistance.

Note that in the model F and A are synergistic since their co-occurrence decreases the *pFR* by an extra –0.09.

The error term is ε, which is the difference between the prediction and the measurement. This error term contains both the *measurement error* on the phenotype measurement and a *model error* (if the underlying model has higher order terms that are not taken into account in the regression model).

Mutation factors for single mutations (M_A, M_B, ..., M_Z) are calculated as follows:

if the mutation is present in the HIV strain, a positive mutation factor is assigned;

if the single mutation is not present, the mutation factor assigned is zero;

if the single mutation is present in a mixture, an averaged positive mutation factor is assigned.

Conveniently, mutation factors range between 0 and 1 where 0 means not present and 1 means present. Values between 0 and 1 means that the mutation is present in a mixture. Accordingly, a positive mutation factor is assigned the value 1.

Mixtures are modelled as causing the average shift of its constituent mutations. Since methods for the quantitation of the precise proportions of mixtures to wild type are expensive and time-consuming, mixture with wild type may conveniently be treated as causing half the *pFR* shift of the resistance mutation (mutation factor = 0.5). However, as the skilled reader will appreciate, a more precise mutation factor may be assigned if the true proportion in the mixture is known.

Mutation factors for double mutations (*MAB etc.*) are calculated as follows;

if both the mutations are present in the HIV strain, a positive mutation factor is assigned (conveniently, the value 1);

if neither of the mutations are present, the mutation factor assigned is zero;

if both mutations are present and one mutation is present in a mixture, an averaged positive mutation factor is assigned (conveniently, 0.5);

if both mutations are present in a mixture, a reduced averaged positive mutation factor is assigned (in this example, 0.25). The factor 0.25 is the product of the M-factors of both the single constituent mutations. This is the result of the assumption that these mixtures are independent of each other. Of course, this is an approximation, since in a real blood sample, the mixtures are not independent of each other. For example, if only 2 viruses were present, virus A (no mutations) for 70% and virus B (mutations 46I and 84V) for 30%, then a mixture would be detected on both positions 46 and 84. If these concentrations were known, it would be possible to fine tune the mutation factor of 0.25. If this information is not available, the best statistical guess is 0.5*0.5: this being the average value that would be measured for the mutation couple being present for a population of samples that have these mixtures on 46 and 84 in all possible concentrations.

Calculation of the resistance coefficient (β_A, β_B, ..., β_Z, β_{AB}) is performed by evaluating the dataset for the drug phenotype reported for each mutation or combination of mutations.

The problem of unwanted correlations has been discussed above. Unwanted correlations are preferably removed according to the method of the invention. A preferred way to do this is to use an algorithm that has been developed by the inventors to track the change in pFR as the effects of individual mutations or combinations of mutations are removed from the dataset. The effect of each mutation or combination of mutations is thus separated out. The methodology follows mutation trajectories towards the global average as the effects of individual mutations or combinations of mutations are removed. The steps are as follows:

1. Calculate average pFR for all mutations with a sufficient count in the database to be significant;

2. Determine the extremes (maximum, minimum), and select the mutation with the pFR furthest away from the global average;

3. Remove all virus strains that have the selected mutation and reiterate from step 1;

4. Stop when the selected mutation in step 2 has an average pFR that approximates to the global average.

In this manner, mutations that do not cause resistance, but which are often present with mutations that do cause resistance will have a higher average pFR (less resistance). Removing the virus strains with a certain resistance causing mutation results in an increase of the average pFR for correlating mutations.

A suitable threshold at which a count in the database becomes sufficiently significant will be apparent to the skilled reader and will be dependent on the database size. For example, thresholds of 5, 10, 15, 20, 25, 30 or more may be suitable. In the examples discussed herein, a threshold of 20 times was used.

By an "average pFR that approximates to the global average" is meant that the average pFR is within a fraction of the standard deviation of the remaining population. A convenient fraction ranges between about 0.3 and 0.5.

A comparison of the change in the global average pFR with the change in the average pFR for selected mutations with increasing iterations of the algorithm is shown in Figure 6a. Figure 6b shows an example, where the average pFR for 71I (unwanted correlation) jumps up as a result of removing from the dataset virus strains that have "71I & 84V" and "48V" mutations.

An alternative, analogous methodology for removing unwanted correlations is as follows; this is an extension of the mutation trajectories algorithm discussed above. The steps of this method are as follows:

1. Calculate correlation coefficient between all mutations (with a sufficient count in the database) and the pFR;
2. Determine the extremes (maximum, minimum), and select the mutation with the highest (absolute value of) correlation coefficient;
3. Calculate a linear model for the pFR with the selected mutation(s) (from step 2, all previous iterations);
4. Take the residue (pFR minus the predicted value from the model);
5. Calculate correlation coefficient between all mutations (with a sufficient count in the database) and the residue;
6. Determine the extremes (maximum, minimum), and select the mutation with the highest (absolute value of) correlation coefficient;
7. Calculate a linear model for the pFR with the selected mutation(s) (from step 6, all previous iterations); and
8. Reiterate from step 4;
9. Stop when the selected mutation in step 7) has a correlation coefficient that approximates to zero.

As with the mutation trajectories algorithm described above, the effect of mutations that do not themselves cause resistance, but which are often present with mutations that do cause resistance, is excluded and thus does not distort the real values.

In further preferred embodiments of the invention, problems of small datasets for particular mutations or combinations of mutations are dealt with by applying the method recursively to the set of virus strains that exhibit those particular mutations or combinations of mutations.

In still further preferred embodiments of the invention, the following additional correlations are taken into account:

- multiple entries of the same virus strain (or virus strains grown from the same stock solution) that cause unwanted correlations;
- censored values in genotype / phenotype database (for example, EC_{50} value = '> 1µM'). These are phenotypes beyond the assay range.

Preferably, censored values are dealt with by attempting to construct a model that is consistent from extrapolations. Censored values are thus modelled by replacing the censored value by a maximum likelihood estimation, assuming knowledge of the standard deviation of the measurement error.

A preferred technique for the generation of a maximum likelihood estimation is as follows:

- Use value V as if the censor was " = ";
- Calculate linear regression model;
- Look at the prediction P from the model:
    - $P < V - 0.798$ (centre of gravity of half Gaussian distribution)
        - Remove value from training data for next iteration
    - $V - 0.798 < P < V$
        - Use $V' = V - 0.798$
    - $V < P$
        - Use V' centre of gravity of tail (<V) of a normal distribution N (P,) as value for next iteration.

Accordingly, for each iteration, when the prediction and measurement contradict, censored values are taken into account. When the prediction and measurement are consistent, censored values are disregarded, on the basis that no further information is provided and their inclusion has no worth.

In one preferred embodiment of this aspect of the invention, the number of calculations necessary in the linear regression analysis may be reduced. The computational power and memory requirement that is currently generally available is insufficient to allow a full second order model to be evaluated for a large dataset, based on all possible single mutations and second order terms, since the number of terms increases quadratically with the number of mutations considered. This number increases with a larger dataset since more rare mutations are in a large database.

In order to reduce the amount of terms, a first order regression may be performed from the list of mutations that occur in the dataset above a threshold number of times. A suitable threshold at which a count in the database becomes sufficiently significant will be apparent to the skilled reader and will be dependent on the database size. For example, thresholds of 5, 10, 15, 20, 25, 30 or more may be suitable. In the examples discussed herein, a threshold of 20 times was used. The significant terms from this first order regression are withheld and the list of these terms is then used to perform a second order regression. In the second order regression only the single mutations and combinations of mutations are used that were found significant in the first order model. Again, a threshold significance will be apparent to the skilled reader – an example is if the probability that the real value of the term is 0, is smaller than 0.001.

For example, a first order regression performed on the matching genotype / phenotype dataset for Indinavir (34,445 measurements) for those mutations that occur at least 20 times results in a first order model that withholds a list of 94 single mutations that are considered significant.

This list is then used as a starting list for a second order regression. It should be noted that it may be advantageous to exclude certain very common mutations from the calculation. 3I is one example. The reason is that a mutation must occur at least a threshold number of times and the inverse also has to be true: the count of viruses *not* having the mutation 3I or the couple *not* 3I and another mutation should also be above the threshold value (e.g. 20). Taking this into account results in excluding 3I from the regression in practice.

In the second order regression, all the single mutations and all couples of mutations from the list are used as potential terms. The significant terms are then withheld by the regression algorithm.

According to a further aspect of the invention, there is provided a method of calculating the quantitative contribution of a mutation pattern to the drug resistance phenotype exhibited by an HIV strain, said method comprising the steps of:

a) obtaining a genetic sequence of said HIV strain,

b) identifying the pattern of mutations in said genetic sequence, wherein said mutations are associated with resistance or susceptibility to drug therapy, and

c) calculating the fold resistance of the HIV strain as compared to the wild type HIV strain by performing a linear regression analysis, whereby the log fold resistance, pFR, is modelled as the sum of all the individual resistance contributions for each of the mutations or combinations of mutations that occur in said HIV strain according to the following equation;

$$pFR = \beta_A M_A + \beta_B M_B + \cdots + \beta_Z M_Z + \varepsilon$$

wherein each individual resistance contribution is calculated by multiplying a mutation factor, M_A , M_B , ..., M_Z, for each mutation or combination of mutations by a resistance coefficient β_A, β_B, ..., β_Z;

wherein the mutation factor assigned reflects the degree to which the mutation or combination of mutations is present in the HIV strain and, if present, to which degree the mutation is present in a mixture;

wherein each resistance coefficient reflects the contribution of the mutation or combination of mutations to the fold resistance exhibited by the strain;

and wherein the error term ε, represents the difference between the modelled prediction and the experimentally determined measurement.

As the skilled reader will appreciate, the fold resistance of the HIV strain may be calculated using any one of the embodiments of the invention referred to above.

In the first step of this method, the genetic sequence of an HIV strain should be obtained. Normally, this will be the genetic sequence of an HIV strain with which a patient is infected, although the sequence may be a theoretical sequence, for example for purposes of *in silico* modelling.

The method may thus be used as a diagnostic method for predicting the fold resistance exhibited by a particular HIV strain with which a patient is infected. According to other preferred embodiments, the method may be used for assessing the efficiency of a patient's therapy or for evaluating or optimizing a therapy. The method may be performed for each drug or combination of drugs currently being administered to the patient so as to obtain a series of drug resistance phenotypes and thus to assess the effect of a plurality of drugs or drug combinations on the predicted fold resistance exhibited by the HIV strain with which the patient is infected.

A "patient" may be any organism, particularly a human or other mammal, suffering from HIV or AIDS or in need or desire of treatment for such disease. A patient includes any mammal and particularly humans of any age or state of development.

To obtain an HIV strain from a patient, a biological sample will need to be obtained from the patient. A "biological sample" may be any material obtained in a direct or indirect way from a patient containing HIV virus. A biological sample may be obtained from, for example, saliva, semen, breast milk, blood, plasma, faeces, urine, tissue samples, mucous samples, cells in cell culture, cells which may be further cultured, etc. Biological samples also include biopsy samples.

The genetic sequence of an HIV strain may be evaluated by a number of suitable means, as will be clear to those of skill in the art. Most suitable will be techniques that allow for specific nucleic acid amplification, such as the polymerase chain reaction (PCR), although other techniques such as restriction fragment length polymorphism (RFLP) analysis will be equally applicable.

Nucleic acid sequencing then allows the analysis of the mutation pattern in a particular nucleic acid sequence, either by classical nucleic sequencing protocols e. g. extension chain termination protocols (Sanger technique; see Sanger F., Nicher., Coulson A. Proc. Nat. Acad. Sci. 1977, 74, 5463-5467) or chain cleavage protocols. Such methods may employ such enzymes as the Klenow fragment of DNA polymerase I, Sequenase (US Biochemical Corp, Cleveland, OH), Taq polymerase (Perkin Elmer), thermostable T7 polymerase (Amersham, Chicago, IL), or combinations of polymerases and proof-reading exonucleases such as those found in the ELONGASE Amplification System marketed by Gibco/BRL (Gaithersburg, MD). Preferably, the sequencing process may be automated using machines such as the Hamilton Micro Lab 2200 (Hamilton, Reno, NV), the Peltier Thermal Cycler (PTC200; MJ Research, Watertown, MA) and the ABI Catalyst and 373 and 377 DNA Sequencers (Perkin Elmer). Particular sequencing methodologies have been developed further by companies such as Visible Genetics. Any of the novel approaches developed for unravelling the sequence of a target nucleic acid, either now or in the future will be perfectly applicable to the analysis of sequence in the present invention (including but not limited to mass spectrometry, MALDI-TOF (matrix assisted laser desorption ionization time of flight spectroscopy) (see Graber J, Smith C., Cantor C. Genet. Anal. 1999, 14, 215-219) chip analysis (hybridization based techniques) (Fodor S P ; Rava R P ; Huang X C ; Pease A C ; Holmes C P ; Adams C L Nature 1993, 364, 555-6) It should be appreciated that nucleic acid sequencing covers both DNA and RNA sequencing.

Once the genetic sequence of the HIV strain is known, the pattern of mutation must be identified in the sequence. The term "mutation" as this is used herein, encompasses both genetic and epigenetic mutations of the genetic sequence of wild type HIV. A genetic mutation includes, but is not limited to, (i) base substitutions: single nucleotide polymorphisms, transitions, transversions, substitutions and (ii) frame shift mutations: insertions, repeats and deletions. Epigenetic mutations include, but are not limited to, alterations of nucleic acids, e. g., methylation of nucleic acids. One example includes (changes in) methylation of cytosine residues in the whole or only part of the genetic sequence. In the present invention, mutations will generally be considered at the level of the amino acid sequence, and comprise, but are not limited to, substitutions, deletions or insertions of amino acids.

The "control sequence" or "wild type" is the reference sequence from which the existence of mutations is based. A control sequence for HIV is HXB2. This viral genome comprises 9718 bp and has an accession number in Genbank at NCBI M38432 or K03455 (gi number : 327742).

Identifying a mutation pattern in a genetic sequence under test thus relates to the identification of mutations in the genetic sequence as compared to a wild type sequence, which lead to a change in nucleic acids or amino acids or which lead to altered expression of the genetic sequence or altered expression of the protein encoded by the genetic sequence or altered expression of the protein under control of said genetic sequence.

A "mutation pattern" comprises at least one mutation influencing sensitivity of HIV to a therapy. As such, a mutation pattern may consist of only one single mutation. Alternatively, a mutation pattern may consist of at least two, at least three, at least four, at least five, at least six, at least seven, at least eight, at least nine or at least ten or more mutations. A mutation pattern is thus a list or combination of mutations or a list of combinations of mutations. A mutation pattern of any particular genetic sequence may be constructed, for example, by comparing the tested genetic sequence against a wild type or control sequence. The existence of a mutation or the existence of one of a group of mutations can then be noted.

One way in which this may be done is by aligning the genetic sequence under test to a wild type sequence noting any differences in the alignment. Typical alignment methods include Smith-Waterman (Smith and Waterman, (1981) J Mol Biol, 147: 195-197), Blast (Altschul *et al.* (1990) J Mol Biol., 215(3): 403-10), FASTA (Pearson & Lipman, (1988) *Proc Natl Acad Sci USA*; 85(8): 2444-8) and, more recently, PSI-BLAST

(Altschul *et al.* (1997) Nucleic Acids Res., 25(17): 3389-402). It may in some circumstances be preferable to generate alignments using a multiple alignment program, such as ClustalW (Thompson *et al.*, 1994, NAR, 22(22), 4673-4680). Other suitable methods will be clear to those of skill in the art (see also "Bioinformatics: A practical guide to the analysis of genes and proteins" Eds. Baxevanis and Ouellette, 1998, John Wiley and Sons, New York). A practical example of multiple sequence alignment is the construction of a phylogenetic tree. A phylogenetic tree visualizes the relationship between different sequences and can be used to predict future events and retrospectively to devise a common origin. This type of analysis can be used to predict a similar drug sensitivity for a sample but also can be used to unravel the origin of different patient sample (i. c. the origin of the viral strain).

In this manner, therefore, the pattern of mutations in the genetic sequence can be identified, wherein said mutations are associated with resistance or susceptibility to drug therapy exhibited by the HIV strain tested. The mutation pattern may influence sensitivity to a specific therapy, e. g., a drug, or a group of therapies. The mutation pattern may, for example, increase and/or decrease resistance of the HIV strain to a therapy. Particular mutations in the mutation pattern, may also, for example, enhance and/or decrease the influence of other mutations present in the genetic sequence that effect sensitivity of the HIV strain to a therapy.

The invention further relates to a diagnostic system as herein described for use in any of the above described methods. An example of such a diagnostic system, for quantitating the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by an HIV strain, comprises:

a) means for obtaining a genetic sequence of said HIV strain;
b) means for identifying the mutation pattern in said genetic sequence as compared to wild type HIV;
c) means for predicting the fold resistance exhibited by the HIV strain using any one of the methods described above.

The means for predicting the fold resistance are preferably computer means.

A still further aspect of the invention relates to a computer apparatus or computer-based system adapted to perform any one of the methods of the invention described above, for example, to quantify the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by HIV, or to calculate the quantitative contribution of a mutation pattern to the drug resistance phenotype exhibited by an HIV strain.

In a preferred embodiment of the invention, said computer apparatus may comprise a processor means incorporating a memory means adapted for storing data; means for inputting data relating to the mutation pattern exhibited by a particular HIV strain; and computer software means stored in said computer memory that is adapted to perform a method according to any one of the embodiments of the invention described above and output a predicted quantified drug resistance phenotype exhibited by an HIV strain possessing said mutation pattern.

A computer system of this aspect of the invention may comprise a central processing unit; an input device for inputting requests; an output device; a memory; and at least one bus connecting the central processing unit, the memory, the input device and the output device. The memory should store a module that is configured so that upon receiving a request to quantify the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by HIV, or to calculate the quantitative contribution of a mutation pattern to the drug resistance phenotype exhibited by an HIV strain, it performs the steps listed in any one of the methods of the invention described above.

In the apparatus and systems of these embodiments of the invention, data may be input by downloading the sequence data from a local site such as a memory or disk drive, or alternatively from a remote site accessed over a network such as the internet. The sequences may be input by keyboard, if required.

The generated results may be output in any convenient format, for example, to a printer, a word processing program, a graphics viewing program or to a screen display device. Other convenient formats will be apparent to the skilled reader.

The means adapted to quantify the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by HIV, or to calculate the quantitative contribution of a mutation pattern to the drug resistance phenotype exhibited by an HIV strain will preferably comprise computer software means. As the skilled reader will appreciate, once the novel and inventive teaching of the invention is appreciated, any number of different computer software means may be designed to implement this teaching.

According to a still further aspect of the invention, there is provided a computer program product for use in conjunction with a computer, said computer program comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising a module that is configured so that upon receiving a request to quantify the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by HIV, or to calculate the quantitative contribution of a mutation pattern to the drug resistance phenotype exhibited by an HIV strain, it performs the steps listed in any one of the methods of the invention described above.

The invention further relates to systems, computer program products, business methods, server side and client side systems and methods for generating, providing, and transmitting the results of the above methods.

The invention will now be described by way of example with particular reference to a specific algorithm that implements the process of the invention. As the skilled reader will appreciate, variations from this specific illustrated embodiment are of course possible without departing from the scope of the invention.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1: Overview of measured /predicted phenotype handling
Figure 2: Phenotype distribution of RTV for matching G/P samples in Virco database
Figure 3: pFR distribution for SQV)
Figure 4a: Distribution of pFR for '48V' mutation on SQV
Figure 4b: Distribution of pFR for '48V' mutation on SQV (expanded)
Figure 5: Removing unwanted correlations
Figure 6a: global mutation trajectories
Figure 6b: mutation trajectories for 71I
Figure 7: Example of genotypes, mutations relative to HBX2
Figure 8: Example of phenotype analysis for RTN
Figure 9: Higher order interaction between mutations 82A and 84V
Figure 10: Illustration of iterative procedure for censored values
Figure 11: Linear regression model identifies mutations included in IAS list. Mutations marked with an * are also identified by a regression on a 5% subset of the data
Figure 12: Linear regression model identifies additional mutations previously described in the literature
Figure 13: Predicted versus measured log(FC)
Figure 14: Comparison between linear regression model and decision trees.

EXAMPLES

Example 1: Methodology

1.1 Introduction

This exercise involved the generation of a list of key mutations for each of the following drugs: Indinavir, Ritonavir, Saquinavir, Nelfinavir, Amprenavir, Lopinavir, Zidovudine, Didanosine, Zalcitabine, Stavudine, Abacavir, Lamivudine, Tenofovir, Nevirapine, Delavirdine and Efavirenz.

The obtained list of key mutations is derived from a linear regression model using single mutations and couples of mutations as independent variables. The dataset used for this analysis is an export of the Virco dataset at 2003/02/01 from the *vircomining* tables. Table 1 shows the matching geno/pheno counts for each drug (*each* phenotype measurement for a genotype counts as one measurement).

Table 1: matching geno/pheno counts

Drug	Count	Drug	Count	Drug	Count
Amprenavir	29,508	Lamivudine	34,395	Delavirdine	32,450
Indinavir	34,445	Abacavir	32,744	Efavirenz	32,601
Lopinavir	7,410	Stavudine	34,420	Nevirapine	34,738
Nelfinavir	34,470	Zalcitabine	34,539		
Ritonavir	34,502	Didanosine	34,227		
Saquinavir	34,543	Tenofovir	14,591		
		Zidovudine	33,575		

1.2 Dataset

The used dataset consists of a set of matching genotype/phenotype measurements with possible multiple phenotype measurements per genotype. The mutations are defined relative to HXB2 at amino acid level. The phenotypes are presented as *pFR* values, which is equal to *- log (FR)*, where *FR* denotes the *Fold Resistance.*

Negative *pFR* values denote resistance and positive values denote hyper-susceptibility. For example, a *pFR* value of -1.0 is equal to 10-fold resistance.

1.3 Linear regression

The derived models are based on the current research on the next generation *Virtual*Phenotype using linear regression models for phenotype prediction. In linear models, the phenotype (*pFR*) prediction is the sum of all individual contributions for each of the mutations in the genotype as in the following equation:

$$pFR = \beta_{10I}M_{10I} + \beta_{10F}M_{10F} + \cdots + \beta_{90M}M_{90M} + \varepsilon$$

The independent variables M_{10I}, M_{10F}, ..., M_{90M} take the values

- 1 if the mutation is present
- 0.5 if the mutation is present in a mixture
- 0 if the mutation is not present

Mixtures are modelled as causing the average shift of its constituent mutations, so mixture with wild type causes half the *pFR* shift of the resistance mutation.

The coefficients β_{10I}, β_{10F}, ..., β_{90M} represent the contribution to the total *pFR* prediction for each single mutation.

The error term is ε, which is the difference between the prediction and the measurement. This error term contains both the ***measurement error*** on the phenotype measurement and a ***model error*** (if the underlying model has higher order terms that are not taken into account in the regression model).

Each M_i represents the presence or absence of the corresponding mutation (absent: 0, present: 1) and each β_i represents the contribution to the pFR change for that specific mutation.

Linear models can contain 2nd order terms (and in general nth order terms) of the form:

$$\beta_{10F90M}M_{10F90M}$$

The independent variable $M10F90M$ represents the co-occurrence of mutations 10F and 90M and the coefficient $\beta 10F90M$ represents the synergy or antagonism between mutations 10F and 90M.

For a 2nd order term, the independent variables M_{10F90M},... take the value:

- 1 if both mutations are present and are not in a mixture
- 0 if one of the mutations is not present
- 0.5 if both mutations are present and one mutation is present in a mixture
- 0.25 if both mutations are present in a mixture

Higher order terms effect for interactions between mutations:

- reversal or antagonism: positive pFR shifts for mutation couples.
- synergy or enhancement: extra negative pFR shift for mutation couple.

For example, consider the following (artificial) model:

Mutation	Coefficient
84V	-0.46
50V	-0.92
54M	-0.64
88S	0.63
90M	-0.16
46L	-0.19
46L & 84V	-0.09

Consider a virus with following mutations: 3I, 46L, 84V and 90M. This virus will have a pFR prediction:

$$pFR=\beta_{46L}.1+\beta_{84V}.1+\beta_{90M}.1+\beta_{46L,84V}.1=-0.9$$

or almost 8-fold resistance. Note that in the model 46L and 84V are synergistic since their co-occurrence decreases the pFR by an extra – 0.09.

Figure 7 (table 1 from poster) shows an example of four different genotypes (mutations relative to HBX2), whilst Figure 8 (Table 2 from poster) shows an example of phenotype analysis for RTV performed according to the method of the invention.

1.4 Model creation

Using our facilities, it was computationally infeasible to calculate a full second order model on all possible mutations and second order terms, since the number of terms increases quadratically with the number of mutations considered.
E.g. for APV:

- Total number of occurring mutations and couples of mutations: 19,074
- mutations and couples with each at least 20 measurements: 4,107

In order to reduce the amount of terms, a first order regression was performed from the list of mutations that occur at least 20 times in the dataset. The significant terms[1] from this regression were withheld and the list of these terms[2,3] was used to perform a second order regression. In the second order regression only the single mutations and couples of mutations are used that were significant in the first order model.

1.5 Example of model creation: *Indinavir*

A first order regression is performed on the matching geno/pheno dataset (34,445 measurements of which 28,480 unique Virco IDs) for those mutations that occur at least 20 times. The resulting first order model withholds a list of 94 single mutations that are considered significant[1]. This list (except 3I) is used as a starting list for a second order regression. In this second order regression, all the single mutations and all couples of mutations from the list are used as potential terms. The significant[1] terms are withheld by the regression algorithm.

[1] A term is called significant if the probability that the real value of the term is 0, is smaller than 0.001

[2] Except mutation 3I for some of the PI's. The reason is that a mutation must occur at least 20 times and the inverse also has to be true: the count of viruses *not* having the mutation 3I or the couple *not* 3I and another mutation should also be at least 20. Taking this into account results in excluding 3I from the regression in practice.

[3] For d4T, the amount of 1st order terms was too high to perform a second order analysis. Only the terms with an absolute value of the coefficient > 0.1 are used in the 2nd order analysis.

1.6 Discussion

Linear regression seeks the model that best fits the underlying data assuming that the underlying data behaves according to a linear model. In that respect, some aspects of this technique have to be taken into account when analysing the results from a regression.

1.6.1 The impact of cross-drug correlation on the significance level of mutations

Correlation between mutations that cause resistance to different drugs, has an impact on the confidence of the coefficient for this mutation. One of the effects is that for non-nucleoside reverse transcriptase inhibitors (NNRTIs) and nucleoside reverse transcriptase inhibitors (NRTIs), some non-relevant mutations for that drug appear as significant (though with a coefficient close to 0), because drug resistance to the drug is correlated with drug resistance to drugs that bind at a different place.

Note that this is only a problem for *interpretation* of the model. For *prediction* of the Fold Resistance, the resulting model remains a good *pFR* predictor.

1.6.2 Effects of second order terms

- Example 1: antagonism

Parameter	pFR shift	Count
...	...	...
82A & 84V	0.43	395
...	...	...
82A	-0.27	4845
...	...	...
84V	-0.26	3531
...	...	...

Second order terms can indicate a *synergy* or an *antagonism*. In the example above, the occurrence of either 82A or 84V cause a resistance shift, but the co-occurrence of both mutations almost completely cancels out the effect of both mutations. In case both mutations are present, the net *pFR* shift is only -0.10, while it is -0.26 or -0.27 if only one of the mutations are present. This is an example of (strong) antagonism.

- Example 2: synergy

Parameter	pFR shift	Count
...	...	...
24I	-0.22	1022
...	...	...
24I & 73S	-0.48	30
...	...	...
73S	-0.20	2216
...	...	...

In this example, 24I and 73S both cause a resistance shift, but their co-occurrence causes a strong extra shift towards resistance. When only one of the mutations is present, the *pFR* shift is -0.20 or -0.22, but the presence of both mutations causes a *pFR* shift of -0.90. 24I and 73S are thus strongly synergistic in this example.

- Example 3: enhancement

Parameter	pFR shift	Count
...	...	...
32I	0	821
...	...	...
32I & 82A	-0.26	516
...	...	...
82A	-0.27	4845
...	...	...

32I by itself does not contribute to resistance, but it increases the resistance for an 82A mutation. 32I enhances the effect of the 82A mutation.

An example of the effects of higher order interactions is shown in Figure 9: (figure 3 from poster).

1.6.3 Highly correlated mutations

Highly correlated mutations (i.e. mutations that almost always co-occur in a strain) can affect the results of a regression analysis. For example, when one of these 2 mutations has an effect on resistance and the other mutation does not (this mutation might for example be highly correlated with the resistance mutation because it affects the replication rate of the virus), the effect can be assigned to either one of the mutations. Unless this is compensated for, the regression model will assign the effect to that mutation that reduces the prediction error the most, which might not always be the mutation that is biologically responsible for the effect. Due to the correlation, it would otherwise not be possible to distinguish between these mutations.

Another effect that occurs due to correlation is when a mutation is highly correlated with a pair of mutations in which the first mutation is present.

Parameter	pFR shift	Count
...	...	
58N	-1.47	108
...	...	
58N & 77L	1.16	106
...	...	
77L	0	471
...	...	

In the above example, 108 samples have a 58N mutation and out of these, 106 samples also have a 77L mutation. The effect of a pure 58N mutation can only be derived from the 2 samples that have 58N and do not have 77L, which leads to higher uncertainty on the estimated *pFR* shift of the 58N mutation. The couple-term '58N & 77L' will compensate for a too low estimation of 58N by having a too high estimation for its *pFR* shift.

Techniques are being developed to deal with these effects. A preferred way to do this is to use an algorithm developed by the inventors to track the change in pFR as the effects of individual mutations or combinations of mutations are removed from the dataset. The effect of each mutation or combination of mutations are thus separated out. The methodology follows mutation trajectories towards the global average as the effects of individual mutations or combinations of mutations are removed. The steps are as follows:

1. Calculate average pFR for all mutations with a sufficient count in the database to be significant;
2. Determine the extremes (maximum, minimum), and select the mutation with the pFR furthest away from the global average;
3. Remove all virus strains that have the selected mutation and reiterate from step 1;
4. Stop when the selected mutation in step 2 has an average pFR that approximates to the global average.

In this manner, mutations that don't cause resistance, but which are often present with mutations that do cause resistance will have a higher average pFR (less resistance). Removing the virus strains with a certain resistance causing mutation results in an increase of the average pFR for correlating mutations.

A suitable threshold at which a count in the database becomes significant is around 20 times.

A comparison of the change in the global average pFR with the change in the average pFR for selected mutations with increasing iterations of the algorithm is shown in Figure 6a. Figure 6b shows an example, where the average pFR for 71I (unwanted correlation) jumps up as a result of removing from the dataset virus strains that have "71I & 84V" and "48V" mutations.

An alternative or even additional methodology for removing unwanted correlations is as follows; this is an extension of the mutation trajectories algorithm discussed above. The steps of this method are as follows:

1. Calculate correlation coefficient between all mutations (with a sufficient count in the database) and the pFR;
2. Determine the extremes (maximum, minimum), and select the mutation with the highest (absolute value of) correlation coefficient;
3. Calculate a linear model for the pFR with the selected mutation(s) (from step 2, all previous iterations);
4. Take the residue (pFR minus the predicted value from the model);
5. Calculate correlation coefficient between all mutations (with a sufficient count in the database) and the residue;
6. Determine the extremes (maximum, minimum), and select the mutation with the highest (absolute value of) correlation coefficient;
7. Calculate a linear model for the pFR with the selected mutation(s) (from step 6, all previous iterations); and
8. Reiterate from step 4;
9. Stop when the selected mutation in step 7) has a correlation coefficient that approximates to zero. By approximates to zero is meant that the correlation coefficient is within a fraction of the standard deviation of the remaining population. A convenient fraction is about 0.4.

As with the mutation trajectories algorithm described above, the effect of mutations that do not themselves cause resistance, but which are often present with mutations that do cause resistance, is excluded and thus does not distort the real values.

1.6.4 Missing second order terms and higher order terms

The regression models were built by first executing a first order regression and then using the list of significant terms from this regression to build a full second order model. However, it is possible that a pair of mutations has an effect while the single mutations do not have an effect by themselves. A first order regression might consider the single mutations as not significant, so these mutations are not used in the second order regression. The couple term for the pair of mutations is therefore not in the final model while this term could be significant. Future regression models can be tuned to overcome this limitation. Alternatively, greater computing power will resolve the need first to perform a first order regression.

The current approach does not involve triplets, quadruples, ... of mutations. It is technically possible to include these terms as well in a linear regression, but higher order terms cause a combinatorial explosion of the number of terms in the regression. This increases the computation time and memory use significantly. Techniques to select a subset of these higher order terms are currently being developed. Expert knowledge can also be a source to select a subset of interesting higher order terms.

1.6.5 Censor values

Censored values occur in the genotype / phenotype database (for example, EC_{50} value = '>1 μM'). These are phenotypes beyond the assay range.

Censored values can be dealt with by attempting to construct a model that is consistent from extrapolations. Censored values are thus modelled by replacing the censored value by a maximum likelihood estimation, assuming knowledge of the standard deviation of the measurement error.

A preferred technique for the generation of a maximum likelihood estimation is as follows:

- Use value V as if the censor was " = ";
- Calculate linear regression model;
- Look at the prediction P from the model:
    - $P < V - 0.798$ (centre of gravity of half Gaussian distribution)
        - Remove value from training data for next iteration
    - $V - 0.798 < P < V$
        - Use $V' = V - 0.798$
    - $V < P$
        - Use V' centre of gravity of tail (<V) of a normal distribution N (P,) as value for next iteration.

Accordingly, for each iteration, when the prediction and measurement contradict, censored values are taken into account. When the prediction and measurement are consistent, censored values are disregarded, on the basis that no further information is provided and their inclusion has no worth.

Figure 10 (figure 4 from poster) illustrates diagrammatically the iterative procedure for censored values.

Example 2: Results

In an initial test, genotypes and corresponding phenotypes determined for ritonavir (RTV) for 28,540 HIV-1 clinical isolates were used. The linear regression analysis identified 20/22 RTV resistance-associated mutations described in the IAS mutation list (all except 10F and 77I) (see Figure 11: (table 4 from poster)). Additional mutations whose effect on RTV susceptibility had been previously described (*e.g.* 73S/T/C, 84A/C and 88D) were also identified (Figure 12: (table 5 from poster)). Overall, 53 single mutations and 96 pairs of mutations were identified as having significant effect on susceptibility to RTV.

The predicted phenotype was compared to the measured phenotype in a leave-one-out cross-validation, demonstrating a root mean square error of 0.31 (logFR) (see Figure 13: (figure 5 from poster). The error rate of the linear modeling method [5.62% (sensitivity=93.0%, specificity=95.4%)], compared favourably to a decision tree-based model [Beerenwinkel, PNAS 99, (2002) 8271-8276] [10.2% (sensitivity=89.8%, specificity=89.7%)] (see Figure 14 (table 3 from poster)).

The robustness of the algorithm as a function of the size of the input dataset was assessed using smaller subsets of data. Nine of 22 IAS resistance-associated mutations for RTV could be identified with subsets ≥ 5% (1600 isolates) of the original data. However, the accuracy of the predicted contribution of the mutations improved with increasing dataset sizes up to 50% of the original database (median standard error of the predicted contributions decreased 50%). Some secondary mutations (e.g. 10R, 32I, 82S) were identified as having a significant contribution to resistance only when the subset size reached a similar 50% level.

We thus conclude that linear regression modelling is a promising new technique for the analysis of drug resistance in HIV-1. It is an attractive tool for identifying primary and secondary resistance-associated mutations for new and existing drugs and for calculating the contribution of mutations and combinations of mutations to resistance. The power of the method is most fully exploited when applied to large datasets of matched genotype/phenotype results.

CLAIMS

1. A method for quantitating the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by HIV, said method comprising the step of performing a linear regression analysis using data from a dataset of matching genotypes and phenotypes,

   wherein the log fold resistance, pFR, of each HIV strain is modelled as the sum of all the individual resistance contributions for each of the mutations or combinations of mutations that occur in HIV according to the following equation;

   $$pFR = \beta_A M_A + \beta_B M_B + \cdots + \beta_Z M_Z + \varepsilon$$

   wherein each individual resistance contribution is calculated by multiplying a mutation factor, M_A, M_B, ..., M_Z, for each mutation or combination of mutations by a resistance coefficient β_A, β_B, ..., β_Z;

   wherein the mutation factor assigned to each mutation or combination of mutations reflects the degree to which that mutation or combination of mutations is present in the HIV strain and, if present, to which degree the mutation is present in a mixture;

   wherein each resistance coefficient reflects the contribution of the mutation or combination of mutations to the fold resistance exhibited by the strain;

   and wherein the error term ε, represents the difference between the modelled prediction and the experimentally determined measurement.

2. A method according to claim 1, wherein for a combination of mutations, the mutation factor M_{AB} represents the co-occurrence of mutations A and B and the coefficient β_{AB} represents the synergy or antagonism between mutations A and B.

3. A method according to any one of the preceding claims, wherein calculation of the resistance coefficient (β_A, β_B, ..., β_Z, β_{AB}) is performed by evaluating the dataset for the drug phenotype reported for each mutation or combination of mutations.

4. A method according to any one of the preceding claims, wherein correlations are removed from the dataset for correlated mutations where not all correlated mutations contribute to the drug resistance phenotype, using an algorithm to track the change in pFR for each mutation as the effects of individual mutations or combinations of mutations are removed from the dataset.

5. A method according to any one of the preceding claims, wherein the algorithm performs the following steps:

   a) calculate average pFR for all mutations with a sufficient count in the database to be significant;

   b) determine the extremes (maximum, minimum), and select the mutation with the pFR furthest away from the global average;

   c) remove all virus strains that have the selected mutation from the dataset and reiterate from step a);

   d) stop when the selected mutation in step b) has an average pFR that approximates to the global average;

   such that removing virus strains with a certain resistance causing mutation results in an increase of the average pFR for correlating mutations, which thus have a higher average pFR.

6. A method according to any one of claims 1-4, wherein the algorithm performs the following steps:

   a) calculate correlation coefficient between all mutations with a sufficient count in the database and the pFR;

   b) determine the extremes (maximum, minimum), and select the mutation with the highest absolute value of correlation coefficient;

   c) calculate a linear model for the pFR with the selected mutation(s) (from step b), all previous iterations);

   d) take the residue;

   e) calculate correlation coefficient between all mutations with a sufficient count in the database and the residue;

   f) determine the extremes (maximum, minimum), and select the mutation with the highest absolute value of correlation coefficient;

   g) calculate a linear model for the pFR with the selected mutation(s) (from step f), all previous iterations); and

h) reiterate from step d;

i) stop when the selected mutation in step h) has a correlation coefficient that approximates to zero.

7. A method according to any one of the preceding claims, wherein multiple entries of the same virus strain or virus strains grown from the same stock solution that cause unwanted correlations are removed from the dataset.

8. A method according to any one of the preceding claims, wherein censored values in the genotype / phenotype database are replaced by a maximum likelihood estimation.

9. A method according to claim 8, wherein for each iteration of the linear regression, the following steps are performed:

a) the censored value [> -X] is initially treated as [= -X];

b) using this value, a linear regression model for predicted pFR is calculated using related values relevant to the pFR of the mutation or combination of mutations;

c) if the calculated model for predicted pFR P is consistent with the censored value, the value is ignored in the next iteration;

d) if the calculated model for predicted pFR P is inconsistent with the censored value, the value is used in the next iteration.

10. A method of identifying a mutation that effects the degree of drug resistance exhibited by an HIV strain using a method according to any one of the preceding claims.

11. A method of calculating the quantitative contribution of a mutation pattern to the drug resistance phenotype exhibited by an HIV strain, said method comprising the steps of:

a) obtaining a genetic sequence of said HIV strain,

b) identifying the pattern of mutations in said genetic sequence, wherein said mutations are associated with resistance or susceptibility to drug therapy, and

c) calculating the fold resistance of the HIV strain as compared to the wild type HIV strain by performing a linear regression analysis, whereby the log fold resistance, pFR, is modelled as the sum of all the individual resistance contributions for each of the mutations or combinations of mutations that occur in said HIV strain according to the following equation;

$$pFR = \beta_A M_A + \beta_B M_B + \cdots + \beta_Z M_Z + \varepsilon$$

wherein each individual resistance contribution is calculated by multiplying a mutation factor, M_A, M_B, ..., M_Z, for each mutation or combination of mutations by a resistance coefficient β_A, β_B, ..., β_Z;

wherein the mutation factor assigned reflects the degree to which the mutation or combination of mutations is present in the HIV strain and, if present, to which degree the mutation is present in a mixture;

wherein each resistance coefficient reflects the contribution of the mutation or combination of mutations to the fold resistance exhibited by the strain;

and wherein the error term ε, represents the difference between the modelled prediction and the experimentally determined measurement.

12. A method according to claim 11, which incorporates a method according to any one of claims 1-9.

13. A diagnostic method for optimising a drug therapy in a patient, comprising performing a method according to any one claims 11-12 for each drug or combination of drugs being considered to obtaining a series of drug resistance phenotypes and therefore assess the effect of the plurality of drugs or drug combinations on the predicted fold resistance exhibited by the HIV strain with which the patient is infected and selecting the drug or drug combination for which the HIV strain is predicted to have the lowest fold resistance.

14. A method according to any one of claims 11-13, wherein the resistance coefficient for each mutation is calculated using a method according to any one of claims 1-9.

15. Use of a method according to any one of claims 1-9 for assessing the efficiency of a patient's therapy or for evaluating or optimizing a therapy.

16. A diagnostic system for quantitating the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by an HIV strain, said system comprising:

a) means for obtaining a genetic sequence of said HIV strain;

b) means for identifying the mutation pattern in said genetic sequence as compared to wild type HIV;

c) means for predicting the fold resistance exhibited by the HIV strain using any one of the methods of claims 1-14.

17. A computer apparatus or computer-based system adapted to perform the method of any one of the claims 1-14.

18. A computer program product for use in conjunction with a computer, said computer program comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising a module that is configured so that upon receiving a request to quantify the individual contribution of a mutation or combination of mutations to the drug resistance phenotype exhibited by HIV, or to calculate the quantitative contribution of a mutation pattern to the drug resistance phenotype exhibited by an HIV strain, it performs a method according to any one of claims 1-14.

ABSTRACT

QUANTITATIVE PREDICTION METHOD

The present invention concerns methods and systems for analysis of drug resistance in HIV-1. More specifically, the invention provides methods for predicting drug resistance by correlating genotypic information with phenotypic profiles. The methods allow the identification of primary and secondary resistance-associated mutations for new and existing drugs and for calculating the contribution of mutations and combinations of mutations to resistance and hypersusceptibility. The invention allows the design, optimization and assessment of the efficiency of a therapeutic regimen based upon the genotype of the disease affecting a patient.

FIG. 1

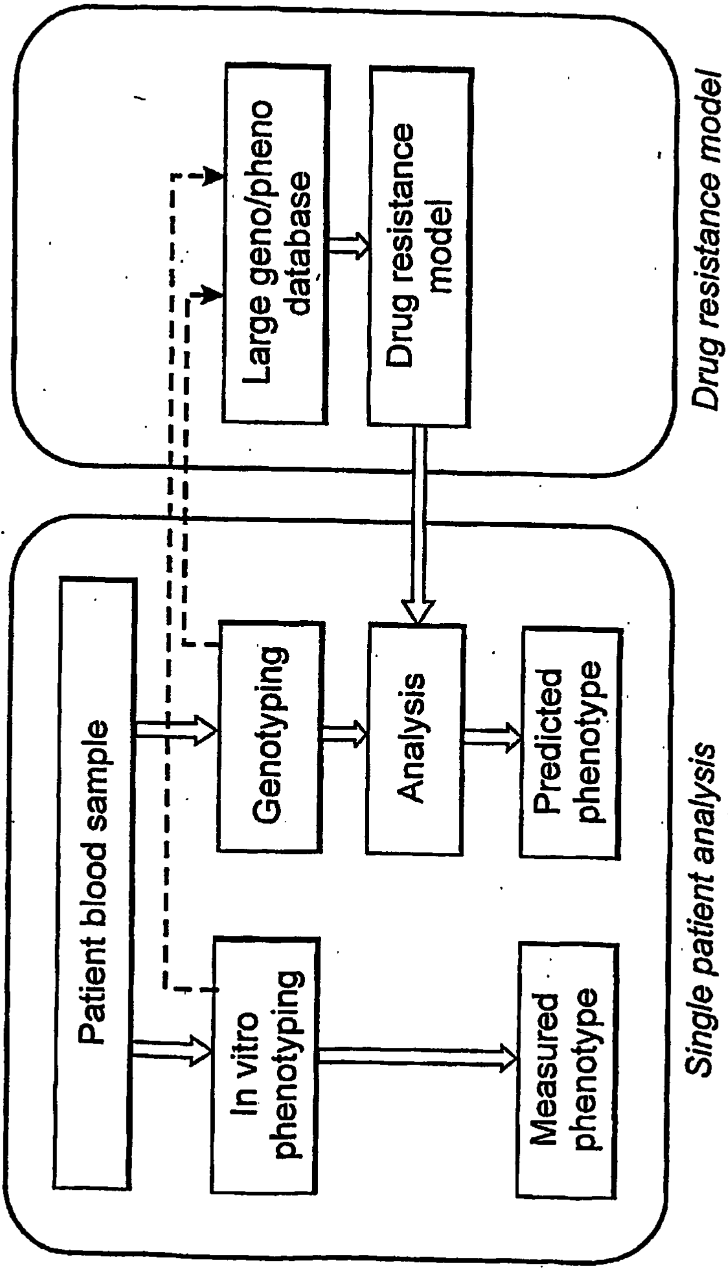

Large geno/pheno database
Drug resistance model
Drug resistance model
Patient blood sample
Genotyping
Analysis
Predicted phenotype
In vitro phenotyping
Measured phenotype
Single patient analysis

FIG. 2

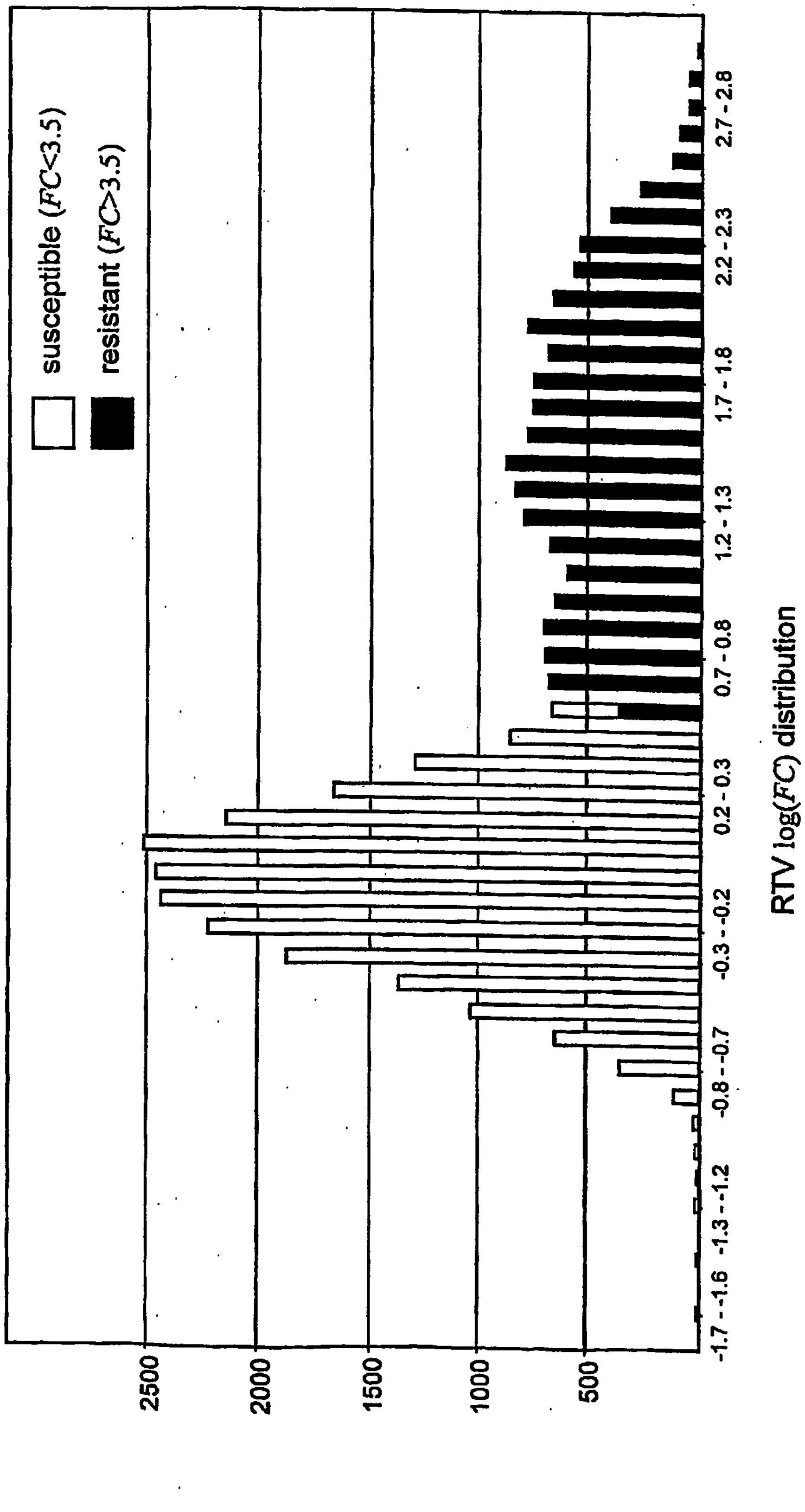

susceptible (FC<3.5)
resistant (FC>3.5)
2500
2000
1500
1000
500
-1.7 - -1.6
-1.3 - -1.2
-0.8 - -0.7
-0.3 - -0.2
0.2 - 0.3
0.7 - 0.8
1.2 - 1.3
1.7 - 1.8
2.2 - 2.3
2.7 - 2.8
RTV log(FC) distribution

FIG. 3

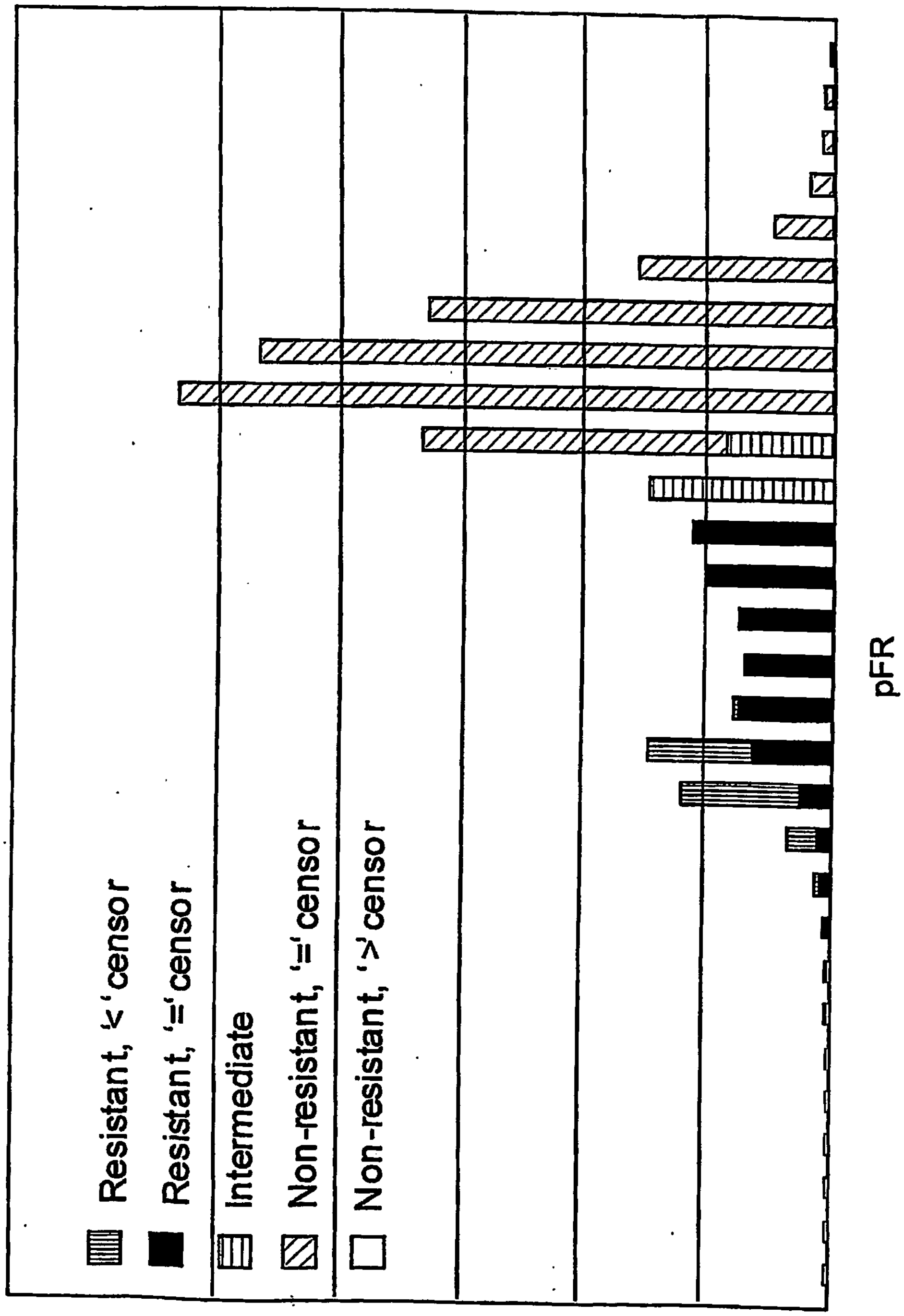
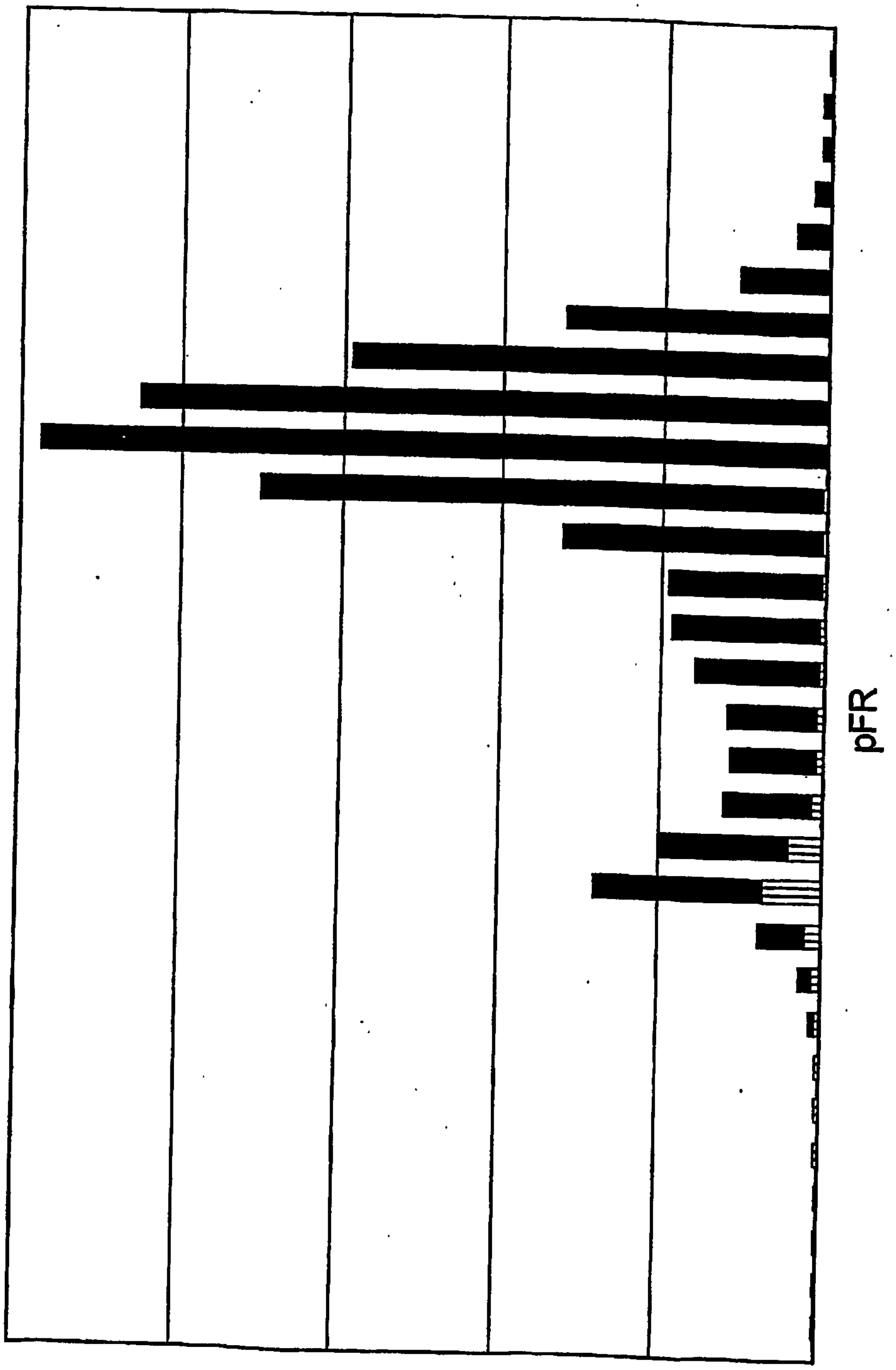

▤ Resistant, '<' censor
■ Resistant, '=' censor
▥ Intermediate
▨ Non-resistant, '=' censor
□ Non-resistant, '>' censor
pFR

FIG. 4a



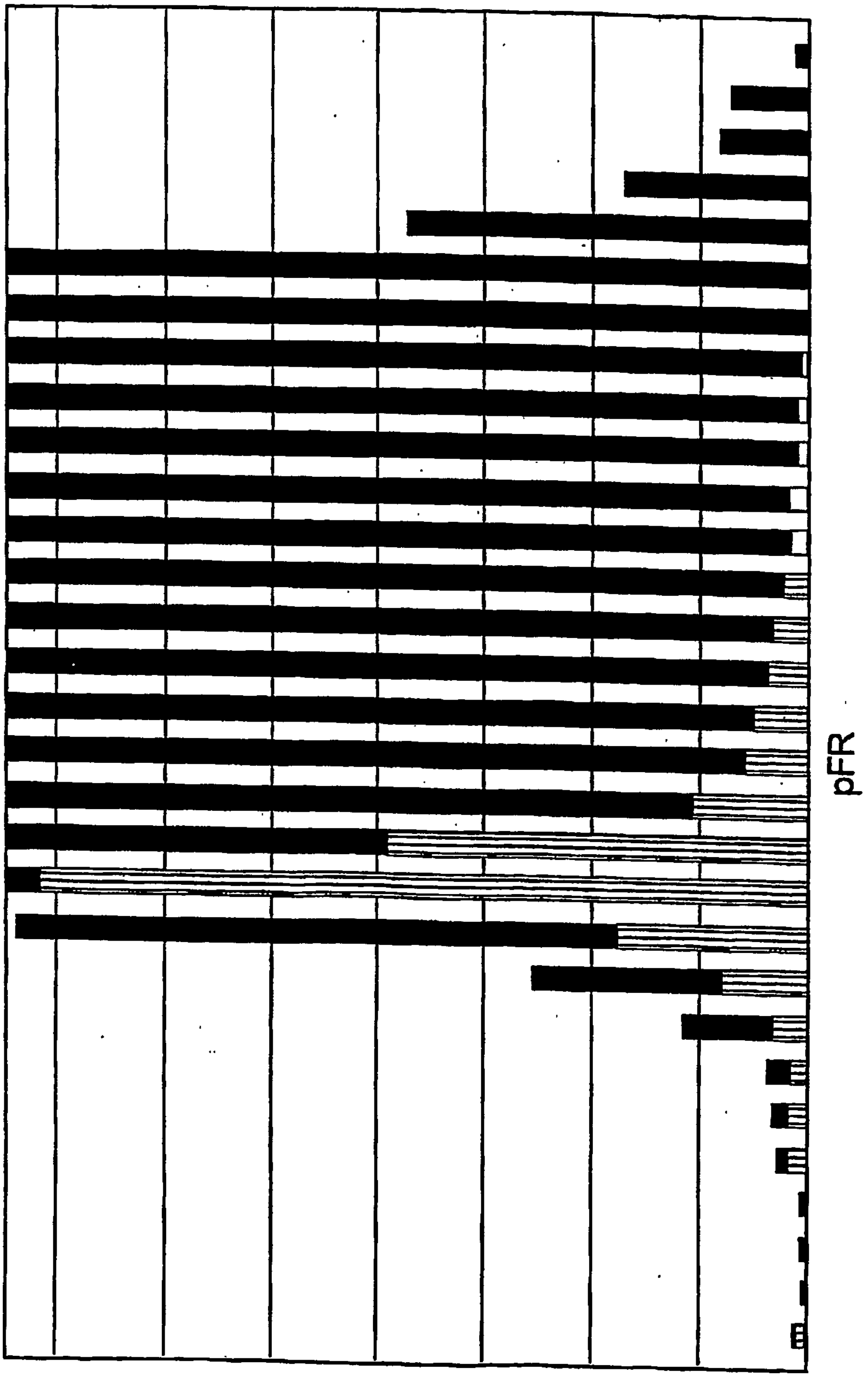


FIG. 4b

FIG. 5

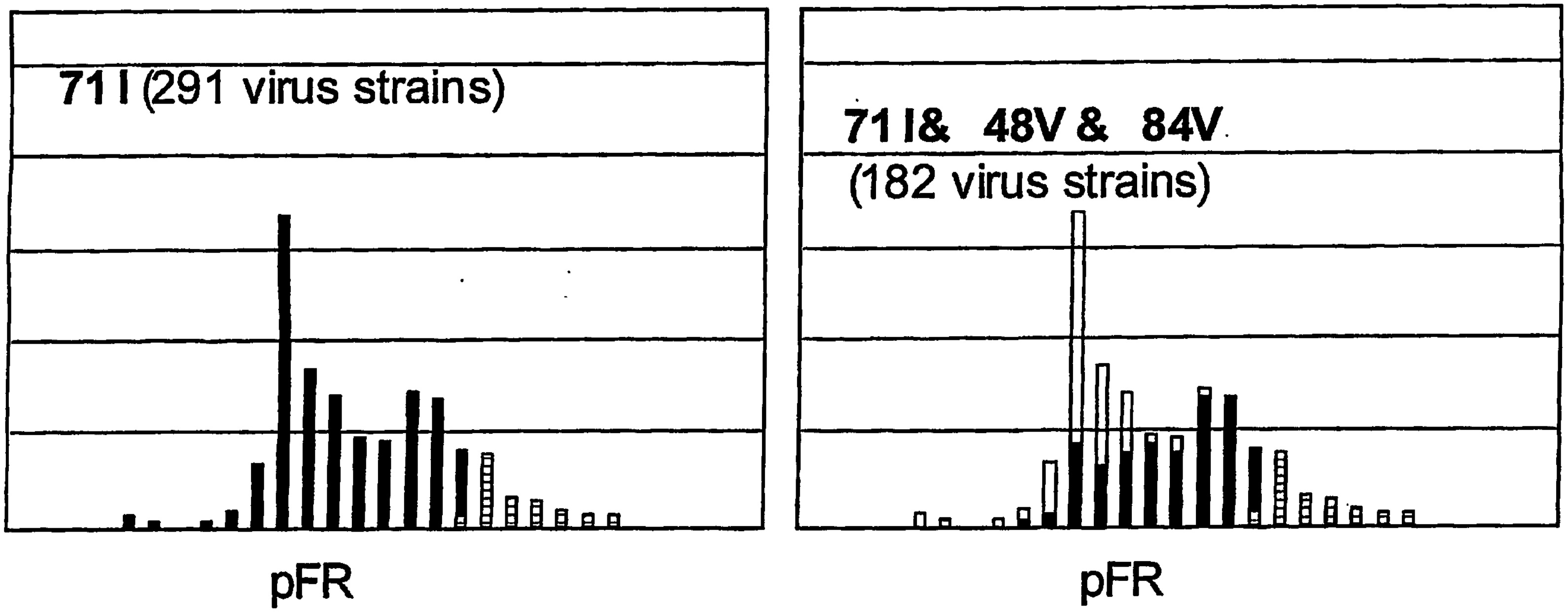


FIG. 6a

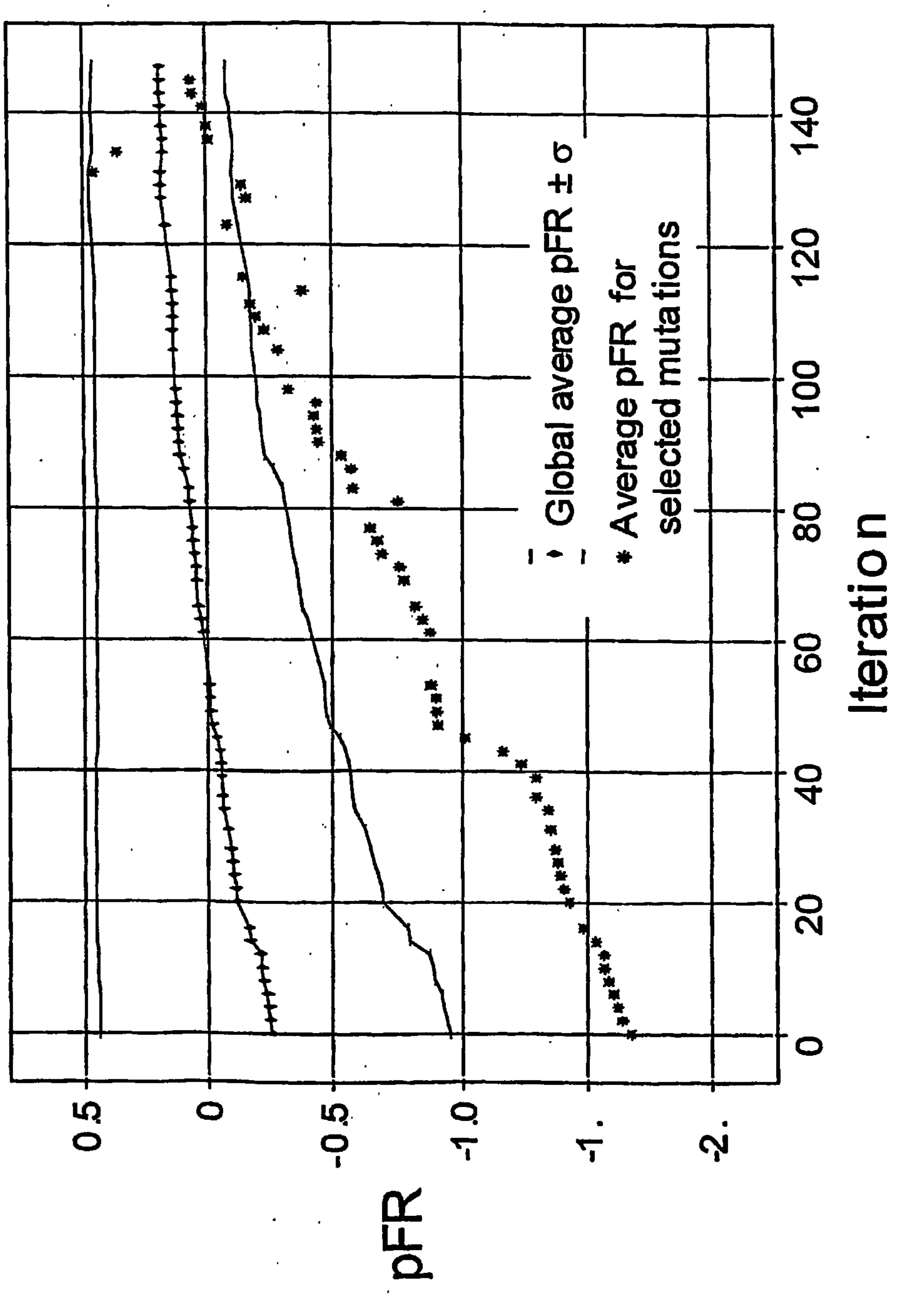

pFR
0.5
0
-0.5
-1.0
-1.
-2.
Iteration
0
20
40
60
80
100
120
140
Global average pFR ± σ
Average pFR for selected mutations

FIG. 6b

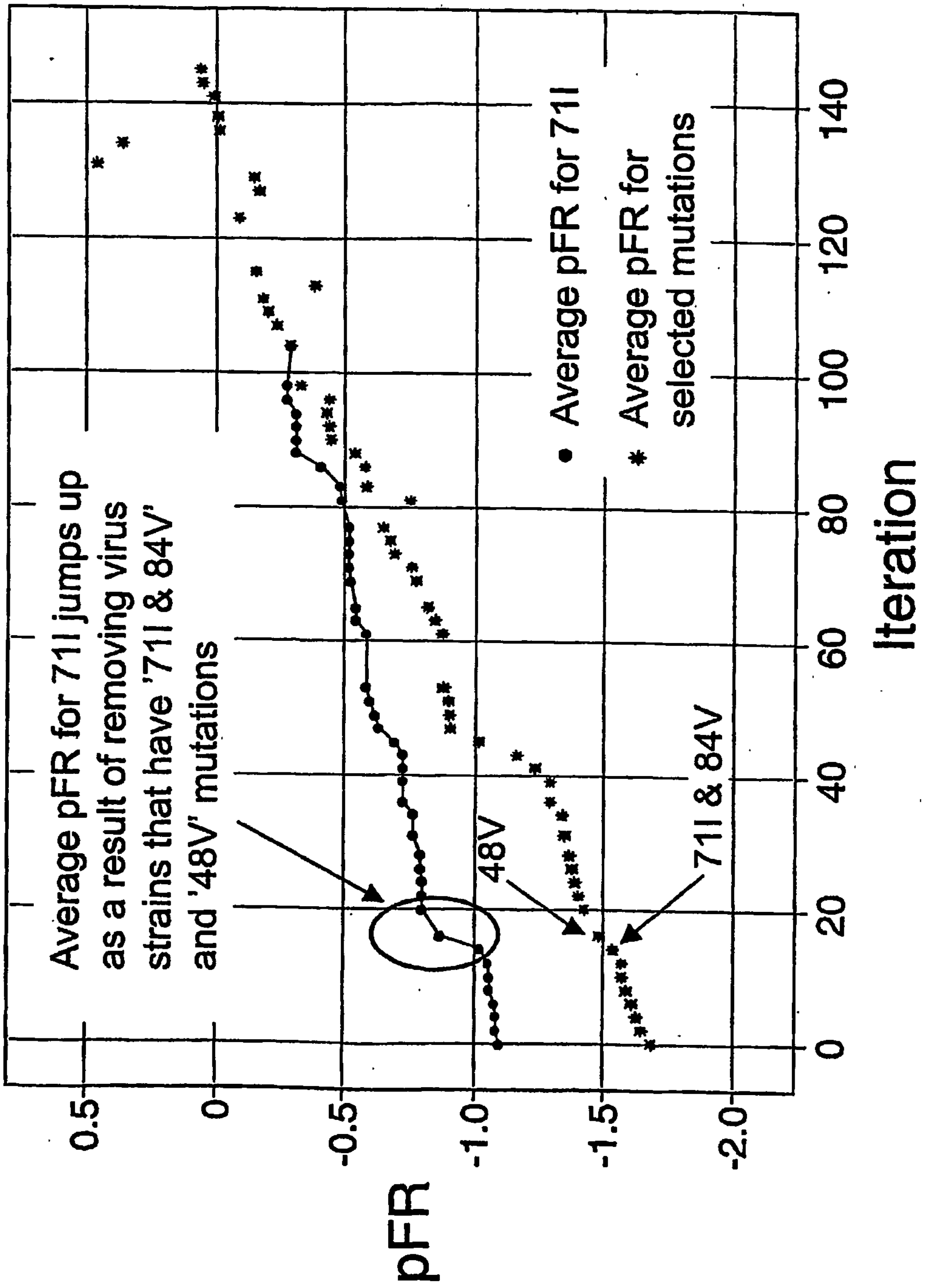

pFR
0.5
0
-0.5
-1.0
-1.5
-2.0
Iteration
0
20
40
60
80
100
120
140
Average pFR for 71I jumps up as a result of removing virus strains that have '71I & 84V' and '48V' mutations
48V
71I & 84V
● Average pFR for 71I
✱ Average pFR for selected mutations

FIG. 7

Sample	3	10	12	15	19	20	24	30	32	36	37	41	46	48	54	62	63	71	74	75	77	82	84	93
V1	I								I/V		N		I				P					A		L
V2	I	I		V		R				I	N	K		V	T		T	V	A			A		L
V3	I		S		I			N			N					V	P			I	I			L
V4	I						I				N		I/M		I/V		P	V				T	I/V	

FIG. 8

Regression model		Virus sample			
Mutation	log(*FC*) shift	V1	V2	V3	V4
24I	0.50				1
30N	-0.39			1	
32I	0.32	½			
46I	0.21	1			½
54T	1.33		1		
54V	0.52				½
82A	0.63	1	1		
82T	0.59				1
84V	0.67				½
Other interactions		*-0.14*	*0.22*	*-0.17*	*-0.02*
predicted log(*FC*)		0.86	2.18	-0.56	1.77
measured log(*FC*)		0.89	2.19	-0.50	1.81

FIG. 9

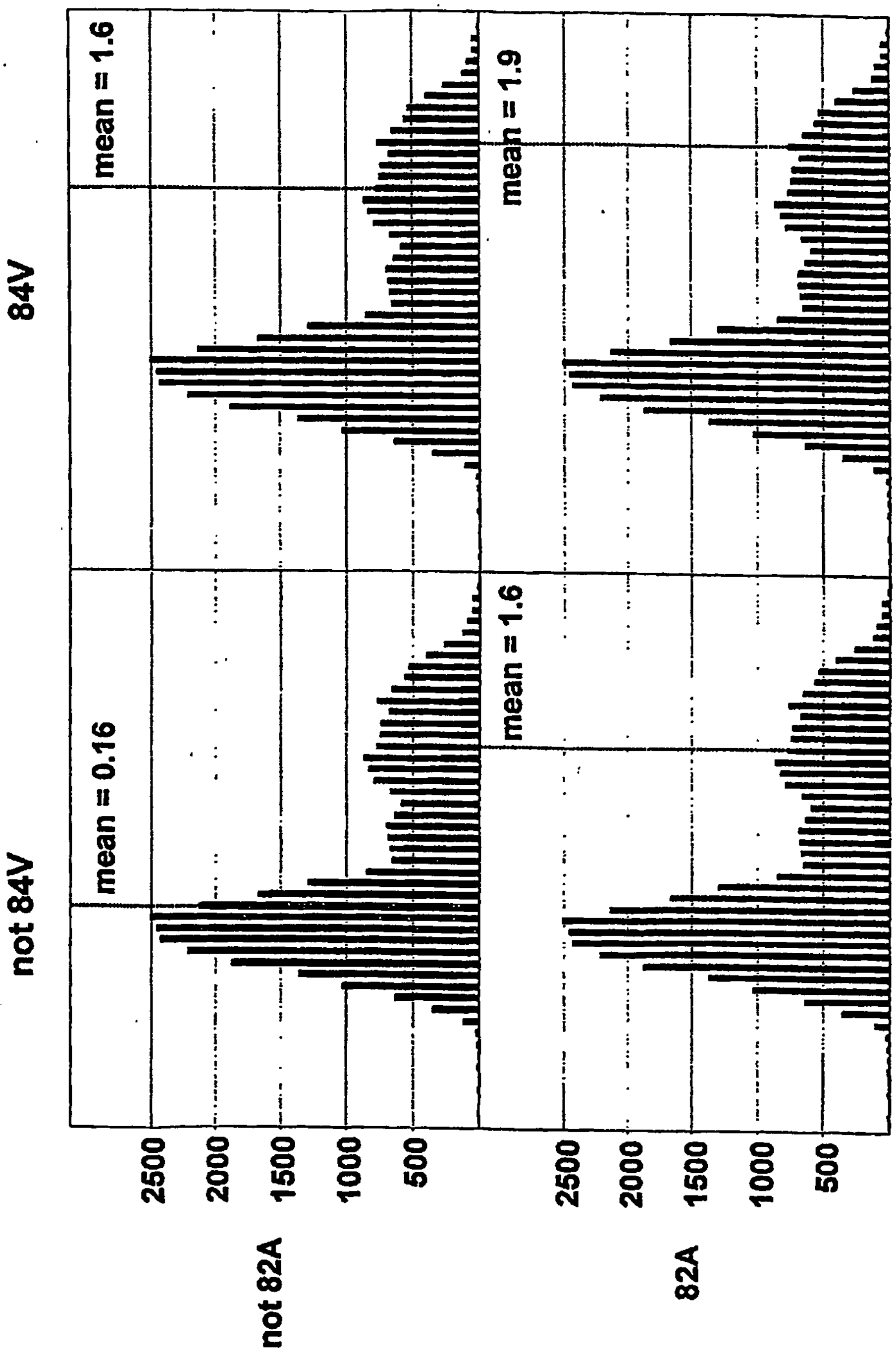

not 84V
84V
not 82A
82A
mean = 0.16
mean = 1.6
mean = 1.6
mean = 1.9
2500
2000
1500
1000
500
2500
2000
1500
1000
500

FIG. 10

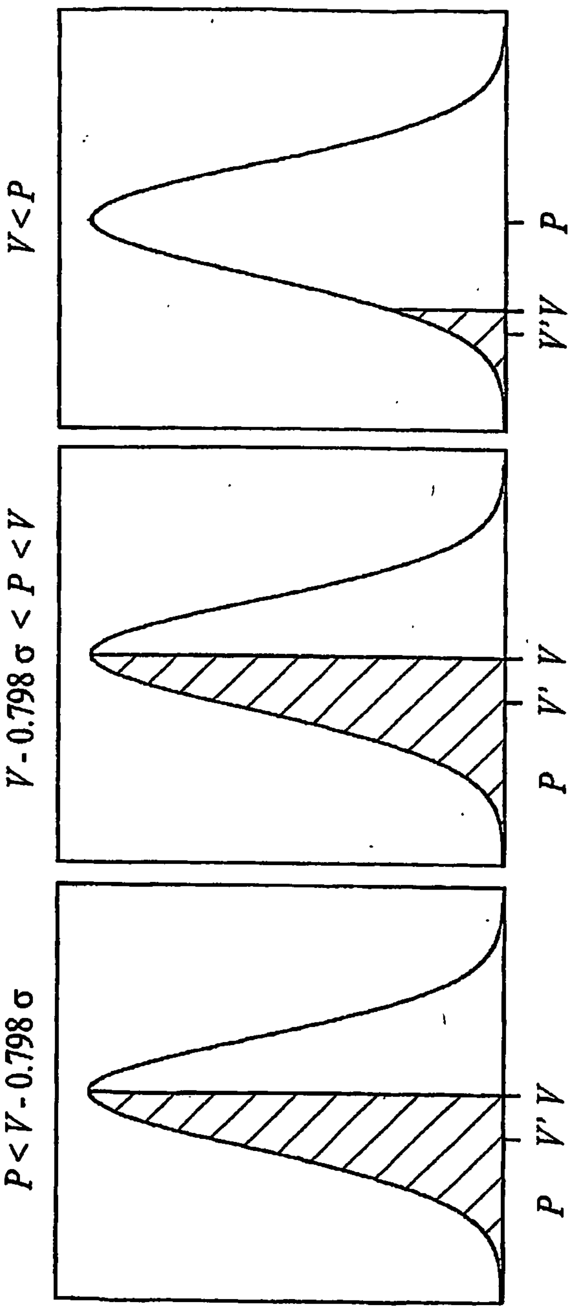


FIG. 11

Mutation	1^{st} order log(*FC*) shift	Prevalence in dataset
10I*	*2^{nd} order terms only*	9,707
10R	0.35	106
10V*	0.15	1,269
20M	*2^{nd} order terms only*	436
20R*	*2^{nd} order terms only*	2,093
32I	0.32	845
33F*	*2^{nd} order terms only*	1,074
36I*	*2^{nd} order terms only*	8,473
46I*	0.21	4,115
46L	0.19	1,745
54L	0.33	367
54V*	0.52	4,553
71T	*2^{nd} order terms only*	2,611
71V	*2^{nd} order terms only*	7,261
82A*	0.63	4,886
82F	0.92	290
82T*	0.59	642
82S	1.17	120
84V*	0.67	3,558
90M*	0.38	9,609

FIG. 12

Mutation	$\log_2(FC)$ shift	Prevalence in dataset
24I	0.50	1,027
30N	-0.39	1,715
54T	1.33	155
73C	0.45	357
73S	0.38	2,224
73T	0.53	559
82M	0.66	33
84A	1.73	70
84C	0.79	67

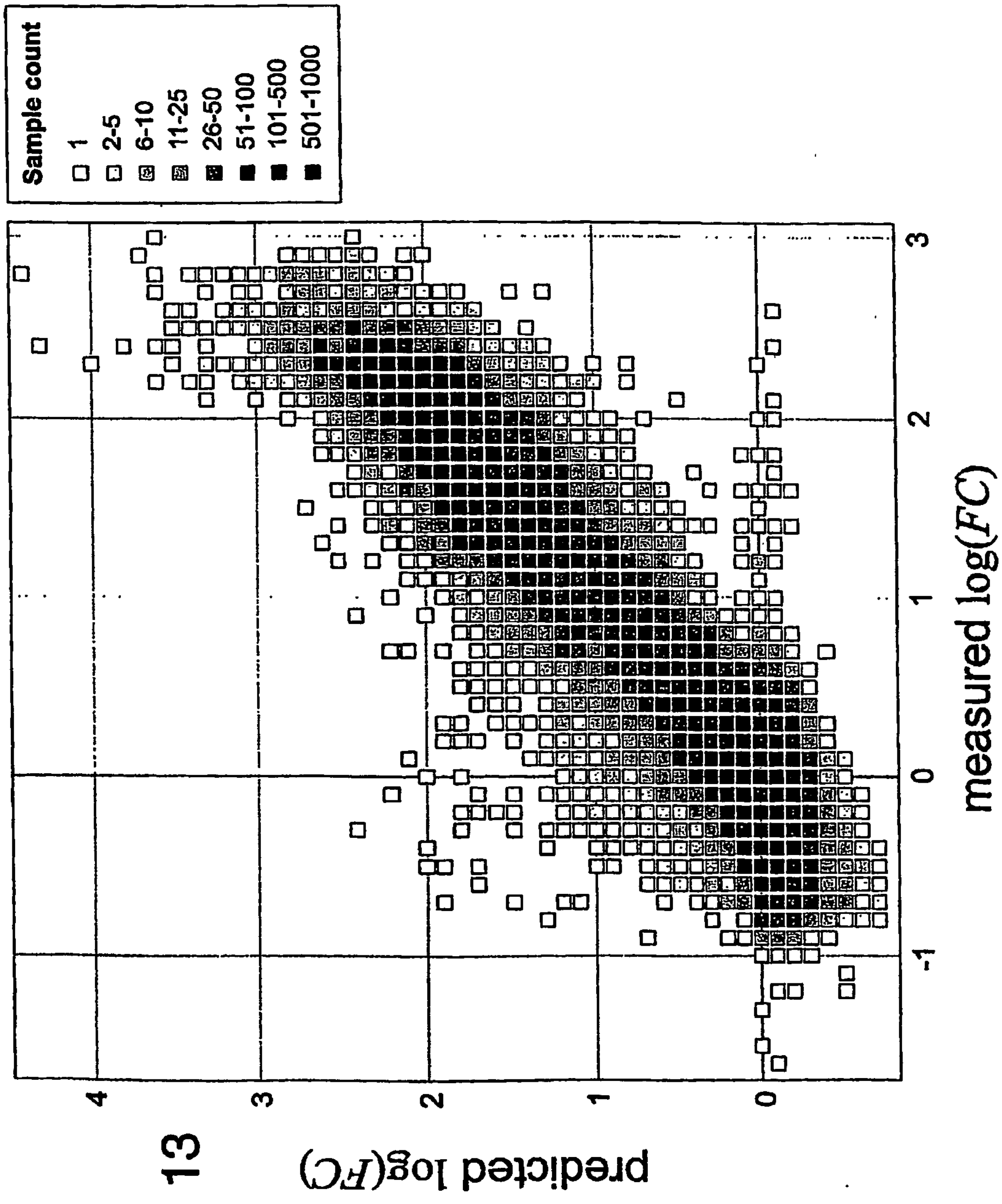

Sample count
□ 1
□ 2-5
▨ 6-10
▨ 11-25
▩ 26-50
■ 51-100
■ 101-500
■ 501-1000
measured log(FC)
-1
0
1
2
3
predicted log(FC)
0
1
2
3
4

FIG. 13

FIG. 14

	Nr. of samples	Resistant fraction (FC>3.5)	Leave-one-out prediction error	Sensitivity	Specificity
Decision tree	469	50.1%	10.2%	89.8%	89.7%
Linear model	469	50.1%	6.4%	92.9%	94.4%
Linear model	34,502	38.3%	5.6%	93.0%	95.4%

- **Regression model identifies** 53 single mutations **and** 96 pairs of mutations **as having a positive or negative contribution to RTV susceptibility**
  20 out of 22 **mutations from** IAS list[1] **are confirmed to be significant by regression model**